PR #21062 完整报告

sgl-project/sglang

Use spec v2 by default

合并时间：2026-04-30 04:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21062

执行摘要

- 一句话：默认启用 spec v2 推测解码
- 推荐动作：该 PR 值得仔细阅读，特别是 server_args.py 中默认逻辑的设计和测试用例的配套调整。展示了如何将一个实验性特性平稳切换为默认，同时保留回退路径。对于使用推测解码的开发者，应了解新的默认行为和降级条件。推荐关注后续是否有针对 topk>1 支持的 PR。

功能与动机

该 PR 旨在将实验性的 spec v2 重叠调度方案设置为默认行为，简化用户配置并释放推测解码的潜在加速收益。此前 spec v2 需要通过环境变量 SGLANG_ENABLE_SPEC_V2=True 手动启用，此次变更使其成为默认，并自动处理不兼容场景的回退。PR 标题和标签表明其专注于 speculative decoding 和 deepseek 模型。

实现拆解

1. 修改服务参数逻辑（python/sglang/srt/server_args.py）：在 _handle_speculative_decoding 方法中，反转默认逻辑。之前默认 spec v1；现在默认启用 spec v2。当 speculative_eagle_topk > 1 或 SGLANG_ENABLE_SPEC_V2 为 False 时自动降级为 spec v1。移除了原先对 topk>1 的硬性 ValueError。同时移除 _handle_model_specific_adjustments 中为 MiMoV2 和 Step3p5 模型强制启用 spec v2 的代码，因为这些模型现在已默认启用。
2. 更改环境变量默认值（python/sglang/srt/environ.py）：将 SGLANG_ENABLE_SPEC_V2 的默认值从 False 改为 True。
3. 清理测试用例中的显式 env override（多个测试文件）：移除那些只在启动服务器时通过 envs.SGLANG_ENABLE_SPEC_V2.override(True) 启用 spec v2 的测试，因为现在默认已启用。受影响文件包括 test_dsa_models_mtp.py、test_qwen3_next_models_mtp.py、test_deepseek_v32_cp_single_node.py、test_deepseek_v32_fp4_mtp_4gpu.py 等。
4. 为依赖 spec v1 的测试添加显式禁用（test/deepep_large.py、test/adaptive_speculative.py）：对于不兼容 spec v2 的后端（如 deepgemm 和自适应推测），使用 with envs.SGLANG_ENABLE_SPEC_V2.override(False) 确保测试在 spec v1 下运行。

5. 新增 Qwen3.5 FP4 测试（test/registered/4-gpu-models/test_qwen35_models.py）：添加 TestQwen35FP4 和 TestQwen35FP4MTP 测试类，覆盖标准推理和 MTP（多 token 预测）路径，默认使用 spec v2。

关键文件：

- python/sglang/srt/server_args.py（模块 服务配置；类别 source；类型 core-logic；符号 _handle_speculative_decoding, _handle_model_specific_adjustments）：核心配置文件，修改了推测解码默认行为逻辑，移除冗余自动启用代码，调整重叠调度条件。
- python/sglang/srt/environ.py（模块 环境配置；类别 source；类型 configuration；符号 SGLANG_ENABLE_SPEC_V2）：修改环境变量 SGLANG_ENABLE_SPEC_V2 的默认值，从 False 改为 True，是默认行为切换的关键。
- test/registered/4-gpu-models/test_qwen35_models.py（模块 模型测试；类别 test；类型 test-coverage；符号 TestQwen35FP4, TestQwen35FP4MTP, TestQwen35FP4MTPV2, test_gsm8k）：新增 Qwen3.5-397B-A17B-NVFP4 模型的 MTP 测试，验证 spec v2 默认下的推理功能。
- test/registered/8-gpu-models/test_dsa_models_mtp.py（模块 模型测试；类别 test；类型 test-coverage）：移除显式的 spec v2 override，测试现在依赖默认值，是清理测试的关键文件。
- test/registered/ep/test_deepep_large.py（模块 模型测试；类别 test；类型 test-coverage）：为 deepgemm 后端显式禁用 spec v2，展示不兼容场景的处理。
- test/registered/spec/eagle/test_adaptive_speculative.py（模块 模型测试；类别 test；类型 test-coverage）：自适应推测暂时不兼容 spec v2，此处显式禁用，保留 spec v1 测试覆盖。

关键符号：server_args.py:_handle_speculative_decoding,
server_args.py:_handle_model_specific_adjustments,
test_qwen35_models.py:TestQwen35FP4.test_gsm8k,
test_qwen35_models.py:TestQwen35FP4MTP.setUpClass,
test_qwen35_models.py:TestQwen35FP4MTP.test_gsm8k

关键源码片段

python/sglang/srt/server_args.py

核心配置文件，修改了推测解码默认行为逻辑，移除冗余自动启用代码，调整重叠调度条件。

```
# 在 _handle_speculative_decoding 方法中，核心逻辑如下
if self.speculative_algorithm in ("EAGLE", "EAGLE3", "STANDALONE"):
    if self.speculative_algorithm == "STANDALONE" and self.enable_dp_attention:
        raise ValueError(
            "Currently standalone speculative decoding does not support dp attention."
        )
    if self.max_running_requests is None:
        self.max_running_requests = 48
        logger.warning(
            "Max running requests is reset to 48 for speculative decoding. "
```

```
            "You can override this by explicitly setting --max-running-requests."
        )

    spec_v1_reason = None
    # topk > 1 时 spec v2 不支持，自动降级
    if (
        self.speculative_eagle_topk is not None
        and self.speculative_eagle_topk > 1
        and not self.disable_overlap_schedule
    ):
        self.disable_overlap_schedule = True
        spec_v1_reason = "spec v2 currently only supports topk = 1"
    # 用户显式关闭了 spec v2 环境变量
    elif (
        not envs.SGLANG_ENABLE_SPEC_V2.get()
        and not self.disable_overlap_schedule
    ):
        self.disable_overlap_schedule = True
        spec_v1_reason = "SGLANG_ENABLE_SPEC_V2=False"

    if self.disable_overlap_schedule:
        logger.warning(
            "Spec v1 is used for eagle/eagle3/standalone speculative decoding because %s.",
            spec_v1_reason or "overlap schedule is disabled",
        )
    else:
        # 默认启用 spec v2
        logger.warning(
            "Spec v2 is enabled by default for eagle/eagle3/standalone speculative decoding."
        )
```

test/registered/4-gpu-models/test_qwen35_models.py

新增 Qwen3.5-397B-A17B-NVFP4 模型的 MTP 测试，验证 spec v2 默认下的推理功能。

```
# TestQwen35FP4MTP 类：测试 spec v2 默认下的 MTP 推理
class TestQwen35FP4MTP(ReasoningTokenUsageMixin, CustomTestCase):
    reasoning_parser_name = "qwen3"

    @classmethod
    def setUpClass(cls):
        cls.model = QWEN35_FP4_MODEL # "nvidia/Qwen3.5-397B-A17B-NVFP4"
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.init_reasoning_token_verifier()
        # 启动服务器时未显式设置 SGLANG_ENABLE_SPEC_V2，依赖默认 True
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
```

```
            "--tp-size", "4",
            "--speculative-algorithm", "NEXTN",
            "--speculative-num-steps", "3",
            "--speculative-eagle-topk", "1",
            "--speculative-num-draft-tokens", "4",
            # ... 其他参数省略
        ],
    )

@classmethod
def tearDownClass(cls):
    kill_process_tree(cls.process.pid)

def test_gsm8k(self):
    args = SimpleNamespace(
        model=self.model,
        eval_name="gsm8k",
        num_examples=200,
        max_tokens=16000,
        temperature=0.6,
        # ...
    )
    metrics = run_eval(args)
    print(f"{metrics=}")
```

评论区精华

未发现公开的 review 讨论。该 PR 由作者 Qiaolin-Yu 自行审批合并，无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 默认行为变更风险：现有部署中如果依赖 spec v1 的特定行为（例如重叠调度在部分后端不稳定），升级后可能遇到性能退化或运行时错误。测试中已对 deepgemm 和自适应推测路径显式禁用 spec v2，但未覆盖所有可能后端。
  2. 静默降级风险：当 topk>1 时，之前会抛出 ValueError；现在自动回退到 spec v1 并仅输出 warning，可能让用户忽略配置不当。
  3. 测试覆盖不充分：大量测试移除了 env override，虽然依赖默认行为，但如果后续默认行为再次变更，测试可能无法检测。新增的 Qwen3.5 测试虽然是正面覆盖，但未测试 topk>1 或禁用 spec v2 的路径。
  4. 环境变量默认值修改（environ.py）可能影响其他依赖该变量判断的应用逻辑。 - 影响：对用户：使用 EAGLE/EAGLE3/STANDALONE 推测解码的用户会自动获得 spec v2 重叠调度，无需手动设置。但之前手动设置 SGLANG_ENABLE_SPEC_V2=True 的用户不受影响，之前未设置的用户行为改变。若用户需要旧行为，必须显式设置环境变量为 False。对系统：服务器启动日志会输出"Spec v2 is enabled by default"或降级原因。重

叠调度默认开启，可能提高解码吞吐，但也可能增加显存占用。对团队：简化了代码维护，不需要在多个模型分支里分散设置 spec v2。测试统一性提高。

- 风险标记：核心默认行为变更，测试覆盖调整，静默降级，兼容性影响

关联脉络

- PR #22416 [Apple Silicon] [MLX] MLX decode partial overlap scheduling for generation (async eval): 同样涉及重叠调度，但针对 MLX 后端，与本 PR 无直接代码耦合，共享重叠调度概念。